

AccuracyTrader: Accuracy-aware Approximate Processing for Low Tail Latency and High Result Accuracy in Cloud Online Services

Rui Han

*Institute Of Computing Technology,
Chinese Academy of Sciences
Beijing, China
hanrui@ict.ac.cn*

Siguang Huang

*School of Software,
Tsinghua University
Beijing, China
huangsg15@mails.tsinghua.edu.cn*

Fei Tang, Fugui Chang, Jianfeng Zhan

*Institute Of Computing Technology,
Chinese Academy of Sciences
Beijing, China
{tangfei, changfugui, zhanjianfeng}@ict.ac.cn*

Abstract—Modern latency-critical online services such as search engines often process requests by consulting large input data spanning massive parallel components. Hence the tail latency of these components determines the service latency. To trade off result accuracy for tail latency reduction, existing techniques use the components responding before a specified deadline to produce approximate results. However, they may skip a large proportion of components when load gets heavier, thus incurring large accuracy losses. This paper presents AccuracyTrader that produces approximate results with small accuracy losses while maintaining low tail latency. AccuracyTrader aggregates information of input data on each component to create a small synopsis, thus enabling all components producing initial results quickly using their synopses. AccuracyTrader also uses synopses to identify the parts of input data most related to arbitrary requests' result accuracy, thus first using these parts to improve the produced results in order to minimize accuracy losses. We evaluated AccuracyTrader using workloads in real services. The results show: (i) AccuracyTrader reduces tail latency by over 40 times with accuracy losses of less than 7% compared to existing exact processing techniques; (ii) when using the same latency, AccuracyTrader reduces accuracy losses by over 13 times comparing to existing approximate processing techniques.

Index Terms—cloud online services; tail latency; result accuracy; synopsis

1. Introduction

Providing quick responsiveness (within 100ms) to user requests is crucial for today's online services such as e-commerce sites and web search engines, as their potential profits are proportional to service latency (request response time) [16], [24], which includes both the request queueing delay and the time of being processed. This paper focuses on a wide class of highly parallel services, in which the

processing of each request needs to consult a large input dataset by parallelizing sub-operations across hundreds or thousands of service components. Each component needs to process a subset of the input dataset to produce a result and hence the *tail latency* (e.g. the 99.9th percentile latency) of these components determines the overall service latency [15], [26]. Example services are: (1) *services using numeric datasets*. At an e-commerce site, a user-based collaborative filtering (CF) recommender system predicts an active user's rating on an unknown item (product) by scanning millions of existing ratings from similar-minded users in a user-item rating matrix [27]. (2) *Services using text datasets*. A web search engine uses an inverted index to organize millions of web pages. For each query, the search engine calculates these web pages' similarity scores to the query words (terms) and ranks the pages in descending order according to their scores.

When delivering services in a cloud platform, service providers usually have limited budgets, namely limited resources, to maintain the quality of service (QoS) requirements of their services. Hence under resource and response time constraints, a wide applied solution is to produce approximate results in request processing in order to trade off result accuracy (correctness) for service latency reduction [12], [14], [23], [24]. For example, in CF-based recommender systems and search engines, the result accuracies are the errors between predicted and actual ratings and the proportion of the actual top k web pages (e.g. the top 10 pages that represent the best answers to the query terms) in the retrieved (returned) top k pages, respectively [14]. As small accuracy losses cannot be evidently perceived and thus are tolerable by service users [12], efficiently and successfully applying such approximate processing mechanism requires reducing component tail latency without incurring large losses in result accuracy.

This task is difficult enough for highly distributed services deployed in a cloud environment, in which service components hosted across different nodes usually have large performance variance. This variance comes from different hardware and software reasons [15] as well as frequently changing performance interference from co-located workloads such as short-running MapReduce jobs [13], [19].

• This paper is accepted as a Regular Paper at the 45th International Conference on Parallel Processing (ICPP-2016).

Furthermore, such performance variance is significantly amplified by request queuing delays when considering service load variations, thus incurring high component tail latency [26]. Existing techniques reduce tail latency by using results only from a part of the components responding before a specified deadline to produce approximate results [15], [23], [24]. However, they do not address issues relating to reducing components' latencies themselves. This means under heavy loads, these techniques have to skip results from a large proportion of slow components to maintain low tail latency. These skipped results may cause large accuracy losses because processing the input data on all components potentially contributes to result accuracy.

In this paper, we propose AccuracyTrader, an approximate request-processing framework for low tail latency and high result accuracy in cloud online services. The basic approach taken by AccuracyTrader is to pre-create a small synopsis to aggregate the information of similar input data points on each component, and then use this synopsis to estimate the correlations between different parts of the input data and arbitrary requests' result accuracy at runtime. AccuracyTrader thus maintains low tail latency by enabling all components producing approximate results quickly using the synopses, while still providing high result accuracy by first using the most accuracy-related input data to improve the produced results. Note that the proposed framework is not intended to replace, but rather complement the existing tail latency reduction techniques based on producing exact results [15], [19], [22], [24]–[26], [28], [29]. AccuracyTrader also differs from traditional techniques that pre-compute structures (e.g. samples or wavelets) of input data based on past query templates and use these structures to answer *certain* types of requests with both accuracy and latency bounds [12]. In contrast, AccuracyTrader needs no prior knowledge about the requests to be processed and it can support arbitrary requests in services.

We have implemented the proposed framework and modified two online services, namely a recommender system [8] and a web search engine [2], to adapt their request processing using AccuracyTrader to study its effectiveness. We first tested the synopsis generation and updating using real-world datasets in both services. The results show that by processing the generated synopses, the parts of input data with higher estimated correlations are indeed more related to different requests' result accuracy. We further compared AccuracyTrader against existing tail latency reduction techniques, using both the synthetic workloads in the recommender system and the realistic search engine workloads derived from the historical user queries of Sogou search engine [11]. The evaluation results show: (i) compared to the request reissue technique based on producing exact results [15], [24], [28], AccuracyTrader reduces the component tail latency by more than 40 times with small accuracy losses of less than 7%; (ii) compared to the partial execution technique based on producing approximate results [15], [23], [24], AccuracyTrader reduces the accuracy loss by more than 13 times when using the same service latency.

The remainder of this paper is organized as follows:

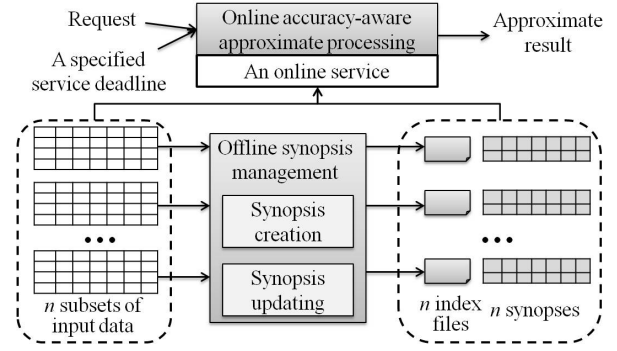


Figure 1. The Overview of AccuracyTrader

Section 2 introduces our approach and Section 3 presents its implementation. Section 4 evaluates the proposed approach. Section 5 discusses the related work, and finally, Sections 6 summarizes the work.

2. AccuracyTrader

In this section, we first present an overview of the AccuracyTrader framework in Section 2.1, following by an explanation of its modules in Sections 2.2 and 2.3.

2.1. Overview

Suppose in an online service, the entire input data is divided into n subsets for parallel processing on n components. AccuracyTrader is presented to enable the accuracy-aware approximate processing on each component using two modules, as shown in Figure 1.

Offline synopsis management. This module is responsible for pre-creating and updating synopses in the offline mode. The module consists of two sub-modules. The *synopsis creation* sub-module organizes each subset of input data using a proper data structure to transform the subset into an index file and a synopsis. The **synopsis** consists of multiple aggregated data points, each aggregates the information of multiple similar data points in the subset. The **index file** records the mapping relationship between each aggregated data point and the original data points aggregated by it. Note that the *synopsis creation* process is only applied once, and the *synopsis updating* sub-module periodically updates the created synopses in an incremental fashion to keep pace with input data changes during service running.

Online accuracy-aware approximate processing. For a request, this module is applied on each component to produce a result using two stages. The first stage produces an initial approximate result using the synopsis, which is sufficiently small (e.g. 100 times smaller than the input data) such that the production process only causes a low latency even when handling heavy loads. By processing the synopsis, this stage also estimates the correlations between different parts of the input data and the request's result accuracy. The second stage iteratively improves the produced result within a specified service deadline. The most

accuracy-related parts are first used in the improvement to minimize the request's accuracy loss.

2.2. Offline Synopsis Management

The basic idea of **synopsis creation** is to group similar data points in a subset of input data and store their aggregated information in a synopsis to preserve data similarity. In AccuracyTrader, R-tree is used in synopsis creation and updating for three reasons. First, in R-tree construction, data points close in feature attributes are allocated to the same node. Second, an R-tree is a depth-balanced tree, which means the nodes at the same depth contain similar numbers of data points and these nodes thus have the same approximation level to the subset. Third, an R-tree is an index structure that supports dynamic insertion and deletion of leaf nodes, thus enabling the incremental updating of an existing synopsis. Based on R-tree, the synopsis creation process has three steps.

Step 1. Dimensionality reduction of the subset. As the R-tree index model works effectively in low-dimensional spaces, this step employs the singular value decomposition (SVD) dimensionality reduction technique to transform the subset into a low-dimensional and dense dataset. SVD can transform a $u \times v$ dataset into a $u \times j$ dataset where j is much smaller than v , while minimizing the difference (distance) between the two datasets. AccuracyTrader uses the incremental SVD [17] whose execution time is independent of the dataset size and hence the transformation process can be completed quickly (within a few seconds) even when dealing with large-scale datasets. Note that the above step works on numeric datasets. For a text dataset such as a collection of web pages, this dataset needs to be transformed into a numeric dataset, in which each data point extracts the feature attribute of its corresponding text data. For example, a web page can be transformed into a numeric data point whose attributes are all the words in the collection of web pages and the value of a attribute is the occurrence number of a word in the web page.

Step 2. Similar data points organization. This step operates on the low-dimensional dataset and groups similar data points in it by constructing an R-tree. In the R-tree, a node including multiple data points corresponds to an *aggregated data point*, and all the nodes at one depth of the tree correspond to the aggregated data points in the *synopsis*. This step outputs an index file by selecting a depth such that it contains a sufficient number of R-tree nodes to enable the fine-grained differentiation of the data points enclosed by different nodes. The number of R-tree nodes at this depth (i.e. the number of aggregated data points in the synopsis) should also be much smaller (e.g. 100 times smaller) than the number of data points in the subset, thus guaranteeing the quick processing of the synopsis.

Step 3. Information aggregation of original data points. According to the index file, the final step obtains each aggregated data point's corresponding *original* data points (without feature reduction) and aggregates their information to generate the synopsis. Depending on the type of dataset,

there are two ways to perform such aggregation. (1) For a numeric dataset, the aggregated information can be the mean of original data points's attribute values. For example, in CF-based recommender systems, suppose an aggregated user (data point) corresponds to a set of a set U of original users, in which a subset $U_i \subseteq U$ of users have rated an item i . The aggregated user's rating on item i is users' average rating on i in set U_i . (2) For a text dataset, the aggregated information can be the merged information of multiple data points. For example, in a search engine, suppose an aggregated web page corresponds to a set of web pages (data points), this page contains all the contents in these pages.

Figure 2 shows an example process of synopsis creation. Step 1 transforms a 12×5 input dataset t into a 12×2 dataset t' . We can see that data points with similar feature attributes (e.g. points d_1 and d_2) in t still have similar attributes in t' . Step 2 organizes the 12 data points in t' by constructing an R-tree, in which similar data points are grouped in the same leaf node. Leaf and non-leaf nodes are then recursively grouped together following the same principle to preserve data similarity. Step 2 selects nodes N_5 and N_6 to generate an index file. Finally, step 3 creates a synopsis consisting of two aggregated data points, each aggregates information of six original data points according the index file.

Motivated by the fact that input data of online services continually changes, **synopsis updating** is designed to periodically update the existing set of synopses. To minimize the overheads in updating, this module detects changes in input data and only updates the synopsis parts influenced by the changes. This updating strategy is built upon the dynamic insertion and deletion of leaf nodes in an R-tree and it considers two situations of input data changes. In the first situation, new input data points are added. This module thus adds new leaf nodes to incorporate these points into the R-tree. In the second situation, the feature attributes or contents of a proportion of existing data points change. This module thus deletes the leaf nodes including these data points and inserts new leaf nodes to represent the changed points. In both situations, synopsis updating identifies the parts of nodes influenced by the newly inserted leaf nodes and updates their corresponding aggregated data points in the synopsis.

2.3. Online Accuracy-aware Approximate Processing

On each component, the steps of accuracy-aware approximate processing are detailed in Algorithm 1. An initial result ar is first produced using the synopsis (line 1). Our current approach uses a sufficiently small synopsis in the production to guarantee a low latency even when handling large service loads. Applying a load-adaptive approach that dynamically selects a synopsis of a different size according to the current load is possible and it is studied in our previous work [18], [20], but it is beyond the scope of this paper.

Estimating the **correlations** between different parts of the input data and the request's result accuracy is the key

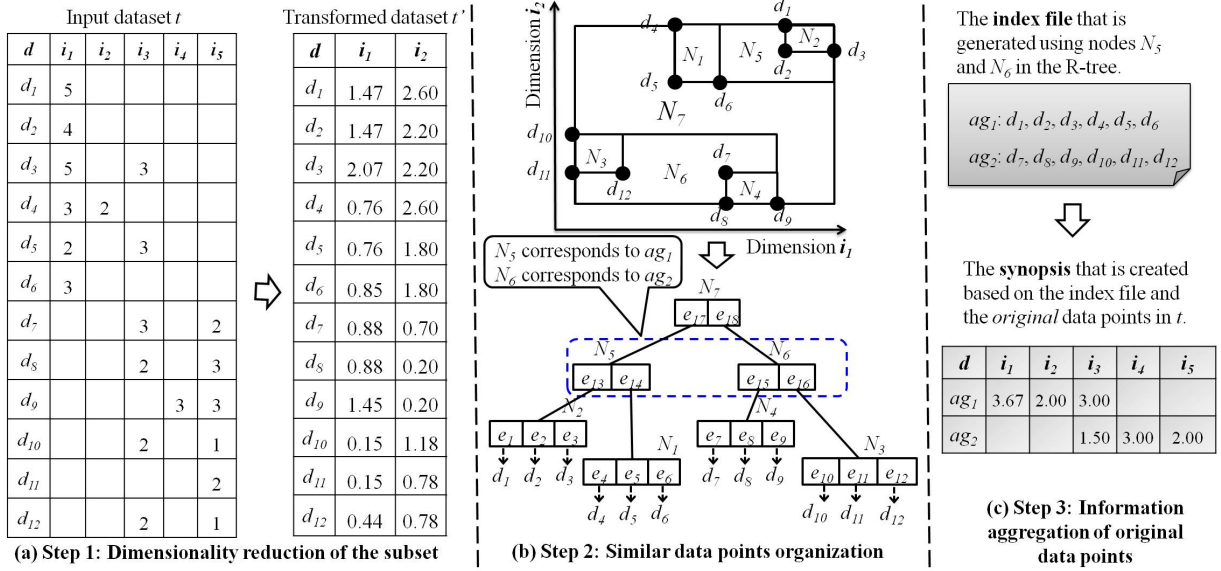


Figure 2. An example of offline synopsis creation

step to enable accuracy-aware request processing and this estimation is based on processing the aggregated data points in synopsis S (line 1). First, processing an aggregated data point ag_i gives an estimation of correlation c_i between this point and the request's result accuracy. For example, in recommender systems and search engines, the correlations are the *weight* between an aggregated user and an active user and the *similarity score* between an aggregated web page and query terms in a request, respectively. In addition, ag_i contains the aggregated information of the original data points in set D_i (i.e. a part of input data) and these points have similar feature attributes. Hence, we assume a linear dependency between c_i and set D_i 's correlation to result accuracy. That is, a higher value of c_i means the accuracy improvement brought by processing the data points in D_i is larger. For example, in search engines, a higher similarity score c_i means the original web pages in set D_i have higher similarity scores on average. Processing the web pages in D_i thus has a higher probability of finding the actual top k web pages and bringing a larger increase in result accuracy.

Based on the estimated correlations, the online module first ranks the aggregated data points (line 2), and then uses the ranking order of each aggregated data point to determine the ranking order of its corresponding set (line 3). Subsequently, the online module sequentially uses the ranked sets to improve result ar (line 4 to 10). The improvement process iteratively executes under two conditions: (1) the elapsed time l_{ela} is smaller than the specified deadline l_{spe} ; (2) the number i of the processed sets is smaller than or equal to i_{max} . The second condition is based on the observation that in some cases, processing the original data points in a proportion of the top ranked sets determines most of the result accuracy. For example, in search engines, the original web pages in the top 40% ranked sets contain over 98% of

the actual top 10 web pages for different requests. Hence, the second condition avoids the unnecessary processing of less accuracy-related data points.

Algorithm 1 Accuracy-aware approximate processing on a component

Require: ag : an aggregated data point;
 c_i : ag_i 's correlation to result accuracy ($1 \leq i \leq m$);
 D_i : the set of original data points represented by ag_i ;
 $S = \{ag_1, ag_2, \dots, ag_m\}$: the synopsis with m points;
 ar : the approximate result;
 l_{spe} : the specified deadline of service latency;
 l_{ela} : the elapsed service time since the request submitting time;
 i_{max} : the maximal number of sets of original data points to be processed.

1. Process S to obtain the initial ar and c_1 to c_m ;
2. Rank the m aggregated data points in descending order according to their correlations to result accuracy;
3. Obtain the ranked sets $\{D'_1, D'_2, \dots, D'_m\}$ according to the ranking orders of aggregated data points;
4. $i = 0$;
5. Obtain the current elapsed time l_{ela} ;
6. **while** ($l_{ela} < l_{spe}$ and $i \leq i_{max}$) **do**
7. Process original data points in D'_i to improve ar ;
8. $i = i + 1$;
9. Obtain the current elapsed time l_{ela} ;
10. **end while**
11. Return ar .

3. Implementation

AccuracyTrader is implemented in Java and it is currently targeted for services running in cloud infrastructures

and Linux environment. Its offline module is implemented based on open source packages of R-tree and SVD (Section 3.1). Its online module is incorporated with two typical parallel online services: a recommender system and a search engine (Section 3.2).

3.1. Offline Synopsis Management Module

AccuracyTrader currently supports R-tree based synopsis creation and updating, which operate on a service’s input data and they are independent of online request processing. First of all, *step 1 of synopsis creation* is implemented based on the incremental SVD method [5]. This step treats the dimensionality reduction process as a gradient descent optimization problem. Suppose a v -dimensional dataset is transformed into a j -dimensional one, the time complexity of this step is $O(j \times i)$, where j is the number of dimensions (e.g. 3) and i is the number of iterations for each dimensionality. *Step 2* is implemented using the standard R-tree package [6]. Given a dataset with k data points, the time complexity of constructing an R-tree is $O(k \times \log k)$. Finally, *step 3* (i.e. information aggregation) is the most computation expensive step, whose time complexity of generating a synopsis using a $k \times v$ dataset is $O(k \times v)$. To accelerate the information aggregation process when dealing with large-scale datasets, we implemented a distributed version of this step running on Spark [3]. This implementation is based on the observation that the information aggregation process typically has a lot of iterative computations (e.g. averaging of feature attributes), which can be significantly accelerated by Spark’s in-memory computing paradigm.

Once the synopsis is generated, the R-tree and the index file are stored and they can be used as the starting point of **synopsis updating**. To minimize the updating overhead, AccuracyTrader uses a low-priority strategy to perform the synopsis updating. On each service component, the synopsis updating sub-module monitors the overall resource utilization and triggers the periodic synopsis updating when the resource is underutilized, thus ensuring little interruption to the running service. This sub-module is implemented to detect newly arrived data or changes in the original input data, dynamically updates the R-tree and the index file, and only re-generates the parts of the synopsis according to the changes in the updated index file.

3.2. Online Accuracy-aware Approximate Processing Module

Incorporating the online module of AccuracyTrader into a service does not require any modification in the request processing algorithm, but controlling the input dataset fed to the algorithm. For each request, the synopsis is first used to produce an initial result and the ranked sets of original data points are then used to improve the result. This implementation is independent of the type of requests to be processed at runtime.

In order to test AccuracyTrader using services of different types, we incorporated its online module into a CF-based

recommender system [8] using numeric input datasets and a Lucene web search engine [2] using text input datasets. We introduce the two services as follows.

CF-based recommender system. In e-commerce sites such as Amazon and eBay, the user-based CF algorithm is a predominant type of techniques applied in many recommender systems [27]. In a CF-based recommendation system, a user-item rating matrix is the input dataset used for storing the user historical ratings (preference scores) for different products (items). For a request from an active user u , the system predicts the u ’s rating on a target item i using two steps. The first step calculates the weight (similarity) between user u and any neighborhood user who has rated the same item i in the matrix. One widely applied weight measure in the CF community is Pearson’s correlation coefficient. The second step generates the prediction of user u ’s rating on item i by taking a weighted average of all ratings of item i from user u ’s neighborhood users.

Lucene web search engine. In today’s Internet, web search engines such as Google, Bing, and Baidu are the most heavily used web services and we study the open source Lucene search engine [2] as an example. At the offline web page collection stage, the web crawler crawls the web pages and builds the inverted index, which includes a vocabulary containing all the words in the crawled web pages. At the online request processing stage, if a query request does not hit the query cache, the search engine scans its index file to search web pages that match the query terms in the request, and then ranks these pages according to their similarity scores to the terms. The service then returns the ranked web pages as the result, in which a small number of top ranked pages (e.g. the top 10 pages) usually stand for the answers to the query terms [14].

We implemented the distributed versions of the above services based on Storm [4] (a real-time distributed processing platform), and incorporated the AccuracyTrader online module. Using AccuracyTrader, the synopsis-based approximate processing operations only causes slightly larger time and space (memory) consumptions than the original service. This is because the synopsis is much smaller than the service input data, and the ranking of the aggregated data points in the synopsis has a polynomial computation complexity depending on the synopsis size.

4. Experimental Evaluation

In this section, we first evaluate the AccuracyTrader offline module using large datasets in real services (Section 4.2). We then compare the AccuracyTrader online module against existing tail latency reduction techniques using different experiment settings (Section 4.3).

4.1. Experimental Settings

Experiment platform. The experiments were conducted on Xen VMs deployed across a cluster with 30 nodes. Each node has two 6-core Intel Xeon E5645 processors, 32GB of DRAM, and eight 1TB 7200RPM SATA disk drives.

Each VM has 2 cores and 4GB memory. The nodes in the cluster are connected through 1Gb ethernet network cards. The operating system of both physical machines and VMs is SUSE Linux Enterprise Server 11 SP1. The Xen, JDK versions are 4.0, 1.7.0, respectively. The enterprise version of Storm, Alibaba JStorm [1], is used. In the JStorm distribution, the versions of JStorm, Python, and Zookeeper are 0.9.6.3, 2.7.6, and 3.4.6, respectively. The version of Hadoop distribution is 1.2.1.

Workloads. We test two service workloads with different request arrival rates based on the implementation of AccuracyTrader on two online services in Section 3.2. We also co-locate both service workloads with Hadoop MapReduce workloads. Two types of MapReduce jobs, namely a CPU-intensive job (WordCount) and an I/O intensive job (Sort), are tested using different input data sizes ranging from 1MB to 10GB. These MapReduce workloads represent a large fraction of short-running and offline batch jobs in the cloud. The MapReduce workloads are generated using BigDataBench-MT [21], a benchmark tool to replay workloads according to real-world traces. The arrival pattern (that is, jobs' submitting time, type, and input data) of MapReduce workloads follows the Facebook production trace provided by Statistical Workload Injector for MapReduce (SWIM) [10], [13].

Compared techniques. The *basic* approach without any tail latency reduction techniques and two state-of-the-art latency reduction techniques are compared: (1) *request reissue* [15], [24], [28]. If some sub-operations of a request have been executed for more than a high percentile of the expected latency for this class of sub-operations, a replica of each straggling sub-operation is sent and only the quicker replica is used. The percentile is set to 95th in our evaluations. (2) *Partial execution* [15], [23], [24]. For each request, this technique only uses a part of sub-operations that complete before a specified deadline to produce its approximate result and skips other sub-operations.

Evaluation metrics. Both performance and accuracy metrics are used to evaluate the online services. The *performance* metric is the 99.9th percentile latency of parallel components for each request. This latency also determines the request's overall service latency. The *accuracy* metric is the percentage of accuracy losses, which denotes the percentage of decreased accuracies in approximate results when comparing to accuracies in exact results that are produced using full computation over the entire input data.

In recommender systems, the *accuracy* is measured by the root-mean-square error (RMSE) [27], which denotes the errors between the predicted and actual values of ratings. Formally, RMSE is a weighted average error that measures the prediction accuracy for all the target items in a test set T :

$$RMSE = \sqrt{\frac{\sum_{i \in T} (p(u, i) - r_{u, i})^2}{n_T}},$$
 where n_T represents the number of items in set T , $p(u, i)$ is the item i 's predicted rating and $r_{u, i}$ is its actual rating. In search engines, the *accuracy* is measured by the proportion of the actual top 10 web pages (i.e. the 10 pages with the highest similarity scores when searching all web pages) in the retrieved top

10 pages [14], [23].

4.2. Evaluation of Offline Synopsis Management

The evaluations in this section first show the overheads of the synopsis generation and updating using the input datasets of both service workloads, and then test of effectiveness of the generated synopses.

Evaluation of overheads of synopsis creation and updating. The CF-based recommender system uses the MovieLens dataset [7] as the input data. In the Lucene search engine, the input dataset is the inverted index created by crawling the Sogou web page collection [9]. The input dataset in both services is divided into 108 subsets. In the recommender system, each subset has approximately 4,000 users, 1000 items, and 0.27 million ratings. In the search engine, each subset has 0.5 million web pages to be searched.

Synopsis creation. We tested the three steps of the synopsis generation on one node. At step 1, a subset is transformed to a 3-dimensional dataset. In SVD transformation, each dimension has 100 iterations. At step 2, the 3-dimensional dataset is organized using an R-tree to generate an index file. In the recommender system, each aggregated user corresponds to an average of 133.01 original users. In the search engine, each aggregated web page corresponds to an average of 42.55 original pages. At step 3, the information of the subset is aggregated to generate a synopsis. For the recommender system and the search engine, a synopsis was created within 30 seconds and 40 minutes, respectively. Note that in the experiments that follow, the above subsets and the generated synopses will be used.

Synopsis updating. We designed two categories of scenarios to evaluate how the synopsis is updated under different changes in the input dataset. In the first category, each subset has a proportion of $i\%$ new data points (users or web pages) being added. In the second category, each subset has a proportion of $i\%$ existing data points being changed. In each scenario, 10 values of i ($i=1, 2, \dots, 10$) were tested on one node. Each test was repeated 10 times for consistency and the average is reported in Figure 3. The evaluation results show: (i) all the updating processes were completed much faster than the synopsis creation processes; (ii) the first category of synopsis updating was completed faster than the second category. This is because in the incremental updating of a synopsis, the first category of scenarios only needs to add new R-tree nodes. In contrast, the second category of scenarios needs to delete existing nodes and add new R-tree nodes, thus leading to longer updating time.

Evaluation of effectiveness of synopses. In the AccuracyTrader framework, an aggregated data point corresponds to multiple original input data points and represents an approximation of them. This evaluation discusses whether the aggregated data points with higher estimated correlations to different requests' result accuracy really correspond to the original data points that are more related to these requests result accuracies.

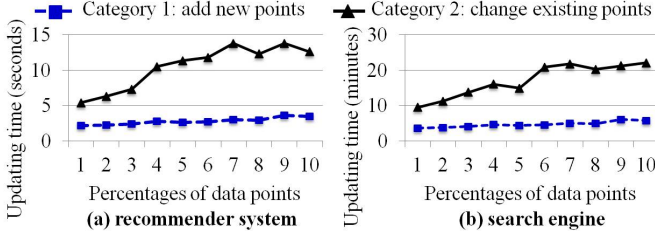


Figure 3. Evaluation of synopsis updating with AccuracyTrader

Evaluation settings. In the recommender system, the data points are users and the request are active users. An *aggregated* data point’s correlation to result accuracy is denoted by the weight (i.e. Pearson’s correlation coefficient) between an active user and an aggregated user. An *original* user is viewed as being highly related to a request’s result accuracy if the weight between the active user and this user is larger than 0.8 or smaller than -0.8 (the weight ranges between -1 and 1). In the search engine, the data points are web pages and the request are queries. An *aggregated* data point’s correlation to result accuracy is denoted by an aggregated web page’s ranking order to a query. An *original* web page is viewed as being highly related to a request’s result accuracy if this page belongs to the query’s actual top 10 web pages. In this evaluation, we randomly selected 1,000 active users (80% of each user’s randomly selected ratings are used in weight calculation) and 1000 queries to represent different requests.

Evaluation results. In Figure 4, the x axis lists the *ranked* aggregated data points divided into 10 sections, and the y axis shows each section’s average percentage of highly related original data points when testing 1000 requests. In the recommender system, the aggregated users are ranked and divided according to their weights to the requests. We can see in Figure 4(a) that the percentage of highly related original users is 95.03% in the first section, this percentage gradually decreases to 22.00% in the last section. In the search engine, the aggregated users are ranked and divided according to their ranking orders. Figure 4(b) shows each section’s average percentage of original web pages that are the actual top 10 web pages for the 1,000 queries. We can see that the first four sections contain 78%, 14.17%, 4.33%, and 1.67% of the actual top 10 web pages respectively, and this percentage is less than 1.17% in the remaining six sections.

Results. Using the synopses, the aggregated data points with higher ranks indeed correspond to parts of input data more related to different requests’ result accuracy.

4.3. Evaluation of Online Approximate Processing

The evaluations in this section compare AccuracyTrader with existing techniques under the same deployment setting.

Deployment settings. We deployed the service (either the recommender system or the search engine) in a cluster of 110 VMs, each hosts a service component. The 110

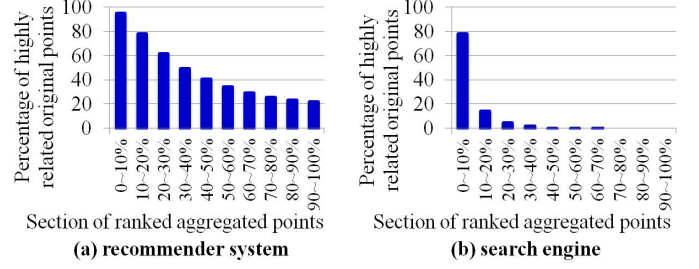


Figure 4. Evaluation of identifying highly related original data points with synopses

service components include one component for accepting and partitioning requests, 108 parallel components for processing the 108 subsets of input data, and one component for composing results to produce responses to end users. The components deployed on each node co-locate with a VM running the Hadoop MapReduce workloads to reflect the changing performance interferences.

Comparison settings. For the *basic* approach and *request reissue* that produce exact results, we compare the performance between them and AccuracyTrader. We also show the accuracy losses of the approximate results produced by AccuracyTrader. For *partial execution* that produces approximate results, we set the same service latency deadline for both techniques and compare their accuracy losses.

Comparison using the synthetic CF-based recommendation workloads. Five request arrival rates, namely 20, 40, 60, 80, and 100 requests/second, were tested. For each test, we randomly selected 1,000 users as the active users from the MovieLens dataset. For each active user, we further randomly selected 20% of items to predict their ratings. AccuracyTrader is set to process as many original data points as possible within the specified deadline because all these points potentially contribute to result accuracy according to Figure 4(a).

Evaluation results. Table 1 shows the tail latencies of the three techniques under different request arrival rates. When the load is light (arrival rate is 20), request reissue has the smallest latency. When load gradually increases, AccuracyTrader provides the lowest tail latencies. Table 2 lists the accuracy losses of partial execution and AccuracyTrader. We can see that in all cases, AccuracyTrader causes accuracy losses of less than 5%, and these accuracy losses are much smaller than those of partial execution.

TABLE 1. COMPARISON OF THE 99.9TH PERCENTILE COMPONENT LATENCY (MS) USING THE CF-BASED RECOMMENDER WORKLOADS

| Request arrival rate | 20 | 40 | 60 | 80 | 100 |
|----------------------|----|-----|-------|--------|--------|
| Basic | 76 | 263 | 48186 | 113496 | 202834 |
| Request reissue | 63 | 213 | 13505 | 27599 | 28981 |
| AccuracyTrader | 87 | 109 | 118 | 122 | 130 |

Comparison using the realistic search engine workloads. In this evaluation, both the query terms and their arrival patterns (that is, queries’ submitting time, arrival rates, and sequences) are derived from a 24-hour user query log

TABLE 2. COMPARISON OF PERCENTAGES OF ACCURACY LOSSES USING THE CF-BASED RECOMMENDER WORKLOADS

| Request arrival rate | 20 | 40 | 60 | 80 | 100 |
|----------------------|------|------|-------|-------|-------|
| Partial execution | 0.26 | 4.50 | 23.39 | 81.48 | 99.56 |
| AccuracyTrader | 0.08 | 0.70 | 1.59 | 2.69 | 4.82 |

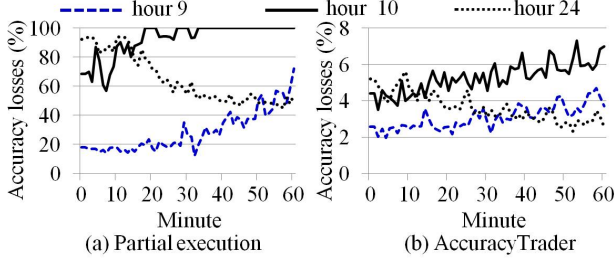


Figure 6. Comparison of percentages of accuracy losses using the search engine workloads of hours 9, 10, and 24

collected by the Sogou search engine [11]. AccuracyTrader is set to process at most the original data points from the top 40% ranked aggregated data points within the specified deadline, because they include over 98.83% of actual top 10 web pages according to Figure 4(b).

We first conduct experiments using *user queries of three typical hours*. As shown in Figures 5(a), (e) and (i), hour 9 (i.e. 8:00 a.m. to 9:00 a.m.), hour 10, and hour 24 represent queries with increasing, steady, and decreasing arrival rates, respectively. We tested each hour separately using 60 sessions, each session lasts 1 minute and the average is reported. Figure 5 demonstrates the fluctuation of tail latencies in three techniques. We can see that the basic approach (Figures 5(b), (f) and (j)) causes the highest tail latencies, which become longer and longer when loads increase because the queueing time of the slowest component continuously increases. In contrast, request reissue (Figures 5(c), (g) and (k)) significantly decreases tail latencies by reducing the latencies of a small proportion of the slowest components. However, this technique still causes much longer tail latencies than those of AccuracyTrader (Figures 5(d), (h) and (l)) when the service is stressed by heavy workloads. In addition, Figure 6 shows that in both approximate processing techniques, the accuracy losses fluctuate with request arrival rates because heavy loads mean less input data can be processed and thus incurring larger losses. AccuracyTrader is affected less by load variations by causing much smaller accuracy losses.

Following the above experimental settings, we extended the comparative experiments using *user queries of 24 hours a day*. Figure 7(a) shows the average request arrival rate of each hour. The average tail latency and accuracy loss at each hour are reported. Figures 7(b), (c), and (d) show the tail latencies of three techniques. Similar to the results in previous evaluations, request reissue has the lowest tail latencies when loads are light (between hour 2 to hour 8), and AccuracyTrader has the smallest latency in other hours. Figure 8 shows AccuracyTrader displays obvious

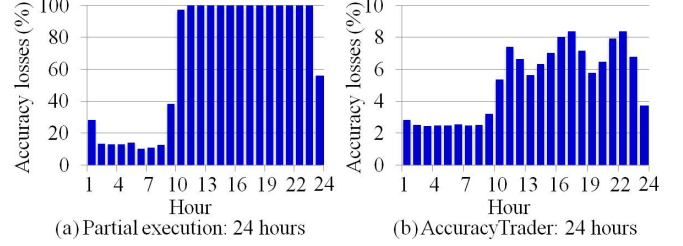


Figure 8. Comparison of percentages of accuracy losses using the search engine workloads of 24 different hours

superiority over partial execution by causing significantly smaller accuracy losses.

Analysis of the compared techniques. *Request reissue* works best when load is light and parallel components have different performances. This technique thus reduces tail latency by reissuing replicas of sub-operations on straggling components to be executed on quick components. Under heavy loads, all components have high latencies because of queueing delays and this fails the reissue mechanism. In contrast, AccuracyTrader achieves consistent low tail latencies by requiring each component completing processing within 100ms. Note that the actual latency is slightly longer than the required one. This is because AccuracyTrader has to process at least the synopsis on each component to produce a result and this processing sometimes causes longer delays than 100ms.

In *partial execution*, each component still performs full computations on the entire input data. Hence under heavy loads, a majority of the components may have longer latencies than the specified deadline. This technique thus has to skip the processing results on these components and incurs large losses in result accuracy. In contrast, although AccuracyTrader only processes a small proportion of input data on each component in order to provide low latencies under heavy loads, the processed data points are the most accuracy-related ones for each request (e.g. only searching 20% of the top-ranked web pages can find over 92% of the actual top 10 web pages), thus only causing small accuracy losses.

Results. Compared to request reissue, AccuracyTrader achieves 133.38 and 42.72 times reductions in the 99.9th percentile latency with small accuracy losses of 1.97% and 6.31% in the evaluations of the recommender system workloads and the search engine workloads, respectively. Using the same service latency, AccuracyTrader achieves 15.12 and 13.85 times reductions in result accuracy losses compared to partial execution in the evaluations of the recommender system workloads and the search engine workloads, respectively.

5. Related Work

Reducing tail latency in highly distributed services has attracted much attentions in recent years [15]. Existing techniques based on producing *exact results* typically fall into

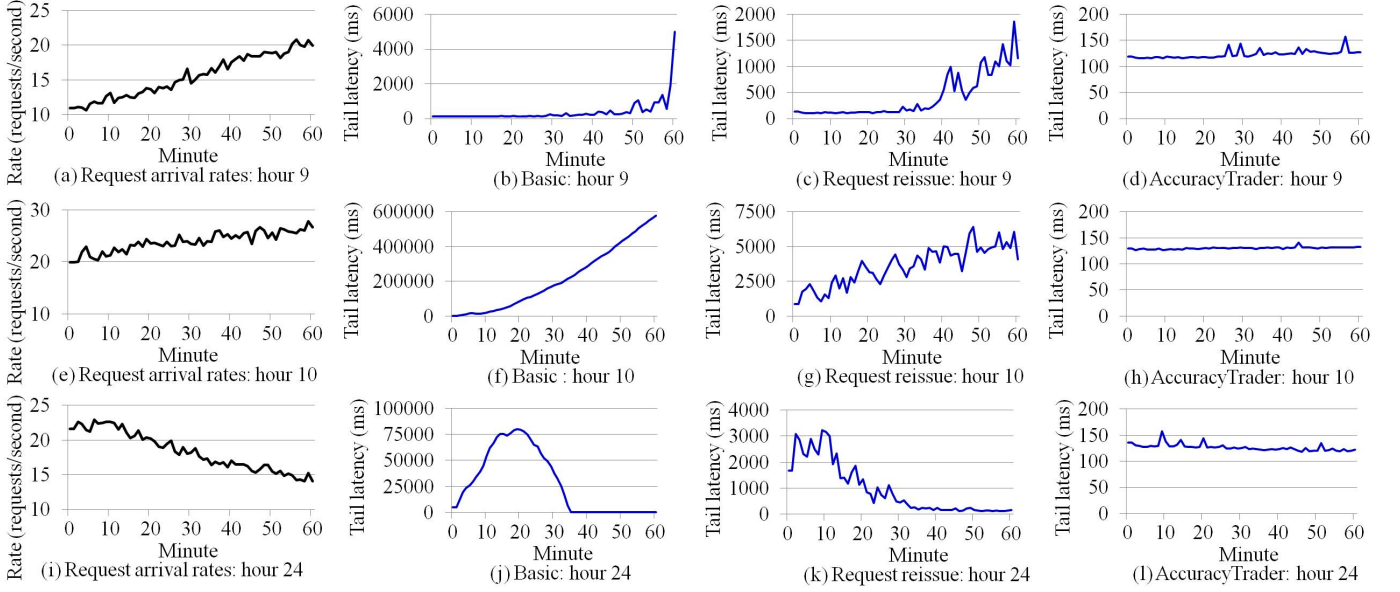


Figure 5. Comparison of the 99.9th percentile component latency (ms) using the search engine workloads of hours 9, 10, and 24

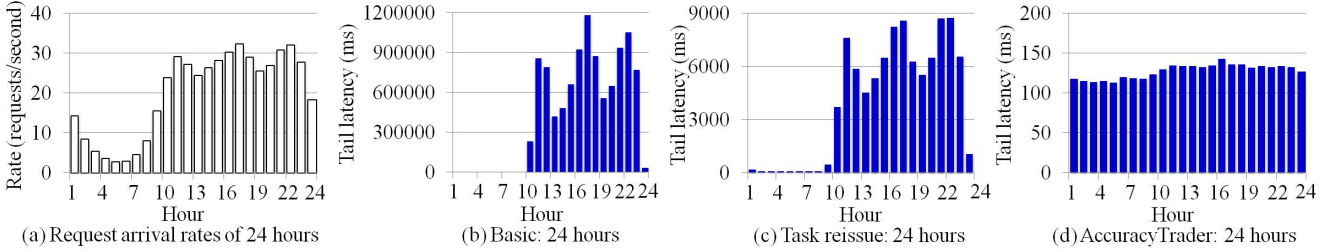


Figure 7. Comparison of the 99.9th percentile component latency (ms) using the search engine workloads of 24 different hours

three categories. The first category uses additional resources to reduce component latency variance, either by increasing the degree of parallelism [25] or by executing redundant requests [29]. The second category modifies the designs of hardware and OS [26] or software systems [22]. The third category mitigates the latencies of straggling components by enforcing dynamic component-node migrations [19] or reissuing requests on these components [15], [24], [28]. Our work forms a complement to these techniques, and we do not explain them in details here. In this section, we discuss related work based on producing *approximate results*.

Approximate processing with accuracy and latency bounds. Based on workload characteristic of past queries, some techniques pre-compute specialized structures (e.g. samples, histograms, or wavelets) of input datasets. Each structure can be used to answer a specific type of query requests with both accuracy and latency bounds [12]. Although these techniques can provide low latency for requests with certain attributes (e.g. the high-frequency terms in search engine), they are impractical to process online services' arbitrary requests, in which the combinations of attributes are unpredictable. Hence these techniques are or-

thogonal to AccuracyTrader, which uses pre-computed synopses that aggregate the entire information of all attributes to support arbitrary requests.

Partial Execution for Tail Latency Reduction. In large-scale online services, the partial execution technique [15], [23], [24] only uses the results from a part of service components responding before a deadline to produce approximate results and skips other components. Although providing low tail latency, all components in this technique still perform exact computations over the entire input data. Hence when loads becomes heavier, a large proportion of components cannot produce results before the deadline and the computations on these components are skipped and wasted. All the skipped results potentially contribute to result accuracy and this technique thus may cause large accuracy losses. In contrast, AccuracyTrader performs computations over sufficiently small synopses to produce quick initial results on all components. Within a specified deadline, it improves the results using the parts of input data most related to result accuracy, thus resulting in high accuracy while maintaining low tail latency despite handling heavy loads.

6. Conclusion

In this paper, we presented AccuracyTrader, an accuracy-aware approximate processing framework for both low tail latency and high result accuracy in cloud online services. AccuracyTrader is based on two key ideas: (1) it aggregates information of similar input data on each component to create a small synopsis, thus enabling all components responding quickly despite handling heavy loads; (2) it estimates the correlations between different parts of the input data and arbitrary requests' result accuracy using the synopsis, thus minimizing accuracy losses by first processing the most accuracy-related input data. Evaluation results using both synthetic and realistic workloads demonstrate the effectiveness of AccuracyTrader at maintaining low tail latency with small accuracy losses.

7. Acknowledgements

We sincerely thank our group members, Junwei Wang, Fengming Ge, and Shulin Zhan, and the anonymous reviewers for their feedback on earlier versions of this manuscript. This work is partly supported by National Natural Science Foundation of China (Grant Nos. 61502451), National High Technology Research and Development Program of China (Grant Nos. Y510091000), and the Key Project of of Guangdong Province, China (Grant Nos. 2015B010108006).

References

- [1] Alibaba jstorm. <https://github.com/alibaba/jstorm>.
- [2] Apache lucene search engine. <http://lucene.apache.org/>.
- [3] Apache spark. <https://spark.apache.org/>.
- [4] Apache storm. <http://storm.apache.org/>.
- [5] Incremental svd method. <http://sifter.org/~simon/journal/20061211.html>.
- [6] Jsi (java spatial index) rtree library. <http://jsi.sourceforge.net/>.
- [7] Movielens 10 million dataset. <http://grouplens.org/datasets/movielens/>.
- [8] Recommender systems. http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/index.html.
- [9] Sogou web pages collection. [Online]. Available: <http://www.sogou.com/labs/dl/t-e.html>.
- [10] Statistical workload injector for mapreduce (swim). <https://github.com/SWIMProjectUCB/SWIM/wiki>.
- [11] User query logs in sogou search engine. <http://www.sogou.com/labs/dl/q-e.html>.
- [12] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, and Ion Stoica. Blinkdb: queries with bounded errors and bounded response times on very large data. In *EuroSys'13*, pages 29–42. ACM, 2013.
- [13] Yanpei Chen, Sara Alspaugh, and Randy Katz. Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads. *Vldb'12*, 5(12):1802–1813, 2012.
- [14] Vinay K Chippa, Srimat T Chakradhar, Kaushik Roy, and Anand Raghunathan. Analysis and characterization of inherent application resilience for approximate computing. In *DAC'13*, page 113. ACM, 2013.
- [15] Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.
- [16] Dan Farber. Google's marissa mayer: speed wins. *ZDNet Between the Lines*, 2006.
- [17] Genevieve Gorrell. Generalized hebbian algorithm for incremental singular value decomposition in natural language processing. In *EACL'06*, 2006.
- [18] Rui Han, Junwei Wang, Fengming Ge, Jose Luis Vazquez-Poletti, and Jianfeng Zhan. Sarp: producing approximate results with small correctness losses for cloud interactive services. In *CF'15*, page 22. ACM, 2015.
- [19] Rui Han, Junwei Wang, Siguang Huang, Chenrong Shao, Shulin Zhan, Jianfeng Zhan, and Jose Luis Vazquez-Poletti. Pcs: Predictive component-level scheduling for reducing tail latency in cloud online services. In *ICPP'15*, pages 490–499. IEEE, 2015.
- [20] Rui Han, Jianfeng Zhan, and Jose Vazquez-Poletti Luis. Sarp: Synopsis-based approximate request processing for low latency and small correctness loss in cloud online services. *International Journal of Parallel Programming*, pages 1–24, 2016.
- [21] Rui Han, Shulin Zhan, Chenrong Shao, Junwei Wang, Jiangtao Xu, Lizy K John, Lu Gang, and Lei Wang. Bigdatabench-mt: A benchmark tool for generating realistic mixed data center workloads. *SoCC'15*, 2015.
- [22] Jun He, Duy Nguyen, Andrea C Arpaci-Dusseau, and Remzi H Arpaci-Dusseau. Reducing file system tail latencies with chopper. In *FAST'15*, pages 119–133. USENIX Association, 2015.
- [23] Yuxiong He, Sameh Elnikety, James Larus, and Chenyu Yan. Zeta: scheduling interactive services with partial execution. In *SoCC'12*, page 12. ACM, 2012.
- [24] Virajith Jalaparti, Peter Bodik, Srikanth Kandula, Ishai Menache, Mikhail Rybalkin, and Chenyu Yan. Speeding up distributed request-response workflows. In *SIGCOMM'13*, pages 219–230. ACM, 2013.
- [25] Myeongjae Jeon, Saehoon Kim, Seung-won Hwang, Yuxiong He, Sameh Elnikety, Alan L Cox, and Scott Rixner. Predictive parallelization: Taming tail latencies in web search. In *SIGIR'14*, pages 253–262. ACM, 2014.
- [26] Jialin Li, Naveen Kr Sharma, Dan RK Ports, and Steven D Gribble. Tales of the tail: Hardware, os, and application-level sources of tail latency. In *SoCC'14*, pages 1–14. ACM, 2014.
- [27] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009.
- [28] Lalith Suresh, Marco Canini, Stefan Schmid, and Anja Feldmann. C3: Cutting tail latency in cloud data stores via adaptive replica selection. In *NSDI'15*, 2015.
- [29] Zhe Wu, Curtis Yu, and Harsha V Madhyastha. Costlo: Cost-effective redundancy for lower latency variance on cloud storage services. In *NSDI'15*, 2015.